# Advances in Parallel Computing and Databases for Digital Pathology in Cancer Research

Siddharth Samsi and Vijay Gadepally
MIT Lincoln Laboratory

*Abstract*—**Over the past decade there have been significant advances in bringing parallel computing and new database management systems to a wider audience. Through a number of efforts such as the National Strategic Computing Initiative (NSCI), there has been a push to merge these "Big Data" and "Scientific Computing" communities to a single computational platform. At the Massachusetts Institute of Technology, Lincoln Laboratory, we have been developing HPC and database technologies to address a number of scientific problems including biomedical processing. In this article, we briefly describe these technologies and how we have used them in the past. We are interested in learning more about the needs of clinical pathologists as we continue to develop these technologies.**

## I. Introduction

Database management systems (DBMSs) have gone through a significant evolution since the original SQL databases. The advent of NewSQL and NoSQL (Not Only SQL) databases has led to the development of new technologies that are well suited for applications beyond traditional database applications such as e-commerce. New DBMS technologies are being developed around supporting traditional scientific workloads such as image processing and correlation. Further, High Performance Computing (HPC) paradigms such as the Message Passing Interface (MPI) have been extended to a number of languages and hardware tools [1] amenable for use in scientific domains. At the Massachusetts Institute of Technology, Lincoln Laboratory, we have been developing tools to bring the power of databases and HPC tools to the common user. In this paper, we present our view of HPC and database technology progression. Further, we present, in detail, one DBMS, SciDB, that is tailored for scientific applications and has the potential for applications in cancer research.

## II. Parallel Computing and Big Data

Parallel computing is the ability to take a given program and split it across multiple processors in order to reduce computation time or resource availability for the application. Very often, advances in parallel processing are directly used for the computational piece of databases such as sorting and indexing datasets. Hadoop [2], for example, supports a parallel file system that forms the basis of a number of DBMSs such as Apache HBase [3]. Parallel computing environments such as pMatlab [4], or bcMPI [5] can significantly reduce the need for deep knowledge of parallel computing. In our previous work, we have applied parallel computing techniques to digital pathology [6]. In these applications, high resolution histopathological images were stored as large TIFF images with JPEG2000 compression and were processed in parallel using MATLAB and the Parallel Computing Toolbox. These approaches require a significant amount of parallel computing development and may not be accessible to the average non-computer scientist interested in developing novel algorithms for disease characterization.

## III. Database Management Systems

Database management systems are very commonly used for indexing and storing large quantities of data. Newer DMBSs go beyond traditional relational systems such as PostGRES [7] to support rapid ingest of data, in-database analytics, hardware accelerated DBMS operations, and data models that more closely resemble the type of data being stored. For example, NoSQL graph databases are tuned to support graph operations and NoSQL key-value databases excel at rapid ingest of unstructured data. Recent NewSQL databases such as MemSQL[8] or Spark [9] leverage main memory (as opposed to disk access) for rapid in-database analytics. One NewSQL database of interest to the scientific community is SciDB which we discuss in detail. Further, the vast proliferation of DBMS technologies has created a new technology selection problem. For this, we have been developing a new type of database called "Polystore" databases that aims to merge the relative advantages of disparate systems.

### A. SciDB

SciDB is a scalable, computational database system that uses an array model for data storage. This array data model stores data in a natural order i.e. the data are stored in the same logical axis system that was used to generate the original data. For example, a 3D volumetric image can be stored in SciDB such that pixels in the image are in the same cartesian space that was used for generating the data and can be accessed by simple array based indexing.

The array-based data model in SciDB also provides support for parallel processing, efficient sparse storage, and in-database linear algebra operations that are well suited for the storage and analysis of biomedical imaging data. SciDB is a full ACID (atomicity, consistency, isolation, durability) DBMS that guarantees repeatability of results across multiple users

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

operating on the same data. One of the unique advantages of SciDB is its ability to perform fast range selects and joins. This capability is achieved by storing data in chunks, in the same order as in the original coordinate system. By storing data in this manner, data that are close to each other can be accessed very quickly by reducing the number of reads necessary to access a given range of data. SciDB also allows a user-settable overlap between chunks of data to speed up applications such as spatial filtering of images for which fewer data reads are necessary for accessing border pixels.

Data stored in SciDB can have multiple attributes per data item. For example, if an RGB color image is stored in SciDB, each entry in the database can have three attributes: Red, Green, Blue pixel values. This is a very powerful concept that can be leveraged for a variety of applications. In the case of weather modeling, data from a variety of sensors gathered at specific geo locations can be stored in the same cell of a SciDB array. Thus, a single query can be used to find geo locations that have similar characteristics such as temperature, wind speed, humidity, etc. This capability has applications in cancer diagnosis and research because of the ability to query and correlate data from a variety of modalities.

### *B. Polystore Systems*

A new approach to designing new systems is to move away from the philosophy that we can design a single system amenable to all possible data types encountered for a particular application [10]. Recently, we have been developing a "Polystore" database system called BigDAWG (short for the Big Data Working Group) that brings together disparate DBMSs. Polystore databases provide a single interface to data stored in disparate systems which are most efficient for a particular dataset. For example, polystore systems can support analytics that require data stored in PostGRES and SciDB simultaneously. We have already applied BigDAWG to medical [11] and genomic data [12] and achieved significant performance and efficiency gains. We have also been developing a software package, D4M (Dynamic Distributed Dimensional Data Model) [13], and related mathematical framework of Associative Arrays [14] to simplify access to polystore systems in a mathematically rigorous manner.

## IV. Application to Cancer Diagnostics

Cancer diagnosis relies on the assessment of stained tissue samples by a pathologists. Depending on the disease, a number of different stains are used to identify disease stage and severity. For example, in Follicular Lymphoma [15], H&E stained tissues are used to count the number of centroblasts and stratify the disease. However, by using adjacent tissue samples stained with CD10, CD20, CD21 antibody researchers may be able to gain valuable insights that are not apparent in isolation. Similarly, in high-throughput imaging, cells are typically labelled with multiple fluorescent proteins that highlight biological phenomenon that are relevant in the drug discovery process. By storing data from multiple channels in the same co-ordinate space, SciDB can be used to develop analytics that can provide new insights into the data. A common approach in the analysis of histopathological images is the conversion of RGB images to other colorspaces [16], [17], [18], [19] such as the La$^*$B or HSV colorspace. By storing multiple colorspace in a single array in SciDB, it is possible to extract regions of interest in the tissue from multiple staining modalities. Additionally, these color space conversions can be performed in-database and stored for future retrieval or use as a pre-processing step. Since SciDB can leverage HPC resources, the data is seamlessly processed in parallel, without the need for extensive re-coding of algorithms for parallel computing.

### *A. Alternative to current approaches*

Analysis of whole slide images is a growing field of research. A common approach to whole slide analysis includes downsampling images, processing only small sections of an image and leveraging HPC systems to perform analysis in parallel. Polystore database and specialized DBMSs like SciDB offer an alternative to these traditional approaches along with the following advantages:

- **Easier data access** Since the image data (2D and 3D) can be stored in it's natural cartesian space, it is easier to develop algorithms that can access the data using natural (x,y,z) co-ordinates rather than reading individual image files to access the data of interest. Other DBMS technologies can be used in conjunction for wider analysis.
- **In-database analytics** Using basic linear algebra approaches, several common processing steps such as color space conversions, smoothing, image math can be computed directly inside the database without the need to extract the image from the database. This pre-processed data can then be stored as attributes of the images in SciDB. Thus, for a given pixel location, it is possible to retrieve not only the original pixel values but also the corresponding pixel values in a different color space.
- **Leverage HPC infrastructure** Since SciDB runs on commodity clusters, it is possible to manage and process massive amounts of data efficiently using commodity hardware. It's capabilities can be extended by adding additional hardware as needed and developing custom analytics that can be performed in-database.
- **Cross-modal analytics with other data** Leveraging tools such as Polystores can greatly simplify the storage and access of related datasets that may be already stored in other systems. For example, time-series data or clinical information can be stored in relational or key-value systems.

## V. Discussion and Conclusions

We believe that there are a number of new technologies that can be used to simplify the development of cancer diagnostics. We have been developing a number of these tools but wish to learn more about what specific capabilities are required by cancer researchers. During this talk, we will describe a number of these advances and learn more about how they can be used to alleviate computational challenges in cancer research.

# References

[1] S. Samsi, V. Gadepally, and A. Krishnamurthy, "Matlab for signal processing on multiprocessors and multicores," *IEEE Signal Processing Magazine*, vol. 27, no. 2, pp. 40–49, 2010.

[2] T. White, *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

[3] L. George, *HBase: the definitive guide*. " O'Reilly Media, Inc.", 2011.

[4] N. T. Bliss and J. Kepner, "'pmatlab parallel matlab library'," *International Journal of High Performance Computing Applications*, vol. 21, no. 3, pp. 336–359, 2007.

[5] D. E. Hudak, N. Ludban, V. Gadepally, and A. Krishnamurthy, "Developing a computational science ide for hpc systems," in *Proceedings of the 3rd international Workshop on Software Engineering for High Performance Computing Applications*. IEEE Computer Society, 2007, p. 5.

[6] S. Samsi, A. K. Krishnamurthy, and M. N. Gurcan, "An efficient computational framework for the analysis of whole slide images: Application to follicular lymphoma immunohistochemistry," *J. Comput. Science*, vol. 3, pp. 269–279, 2012.

[7] M. Stonebraker and L. A. Rowe, *The design of Postgres*. ACM, 1986, vol. 15, no. 2.

[8] N. Shamgunov, "The memsql in-memory database system." in *IMDM@ VLDB*, 2014.

[9] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets." *HotCloud*, vol. 10, pp. 10–10, 2010.

[10] M. Stonebraker and U. Cetintemel, "" one size fits all": an idea whose time has come and gone," in *21st International Conference on Data Engineering (ICDE'05)*. IEEE, 2005, pp. 2–11.

[11] A. Elmore, J. Duggan, M. Stonebraker, M. Balazinska, U. Cetintemel, V. Gadepally, J. Heer, B. Howe, J. Kepner, T. Kraska *et al.*, "A demonstration of the bigdawg polystore system," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1908–1911, 2015.

[12] V. Gadepally and Chisholm Lab. Genomics data, analytics and the future of climate change. [Online]. Available: http://istc-bigdata.org/index.php/genomics-data-analytics-and-the-future-of-climate-change/

[13] V. Gadepally, J. Kepner, W. Arcand, D. Bestor, B. Bergeron, C. Byun, L. Edwards, M. Hubbell, P. Michaleas, J. Mullen *et al.*, "D4m: Bringing associative arrays to database engines," in *High Performance Extreme Computing Conference (HPEC), 2015 IEEE*. IEEE, 2015, pp. 1–6.

[14] J. Kepner and V. Gadepally, "Adjacency matrices, incidence matrices, database schemas, and associative arrays," in *International Parallel & Distributed Processing Symposium Workshops (IPDPSW). IEEE*, 2014.

[15] E. S. Jaffe, N. L. Harris, H. Stein, and J. W. Vardiman, "World health organization classification of tumours of haematopoietic and lymphoid tissues," *IARC Press*, 2001.

[16] S. Samsi, G. Lozanski, A. Shanarah, A. K. Krishanmurthy, and M. N. Gurcan, "Detection of follicles from ihc-stained slides of follicular lymphoma using iterative watershed," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2609–2612, Oct 2010.

[17] S. Samsi, W. N. Jarjour, and A. Krishnamurthy, "Glomeruli segmentation in h amp;e stained tissue using perceptual organization," in *Signal Processing in Medicine and Biology Symposium (SPMB), 2012 IEEE*, Dec 2012, pp. 1–5.

[18] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunyé, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images: a comparative study," *Journal of Digital Imaging*, vol. 29, no. 4, pp. 496–506, 2016.

[19] K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 951 – 961, 2012, special Issue on Awards from {ICPR} 2010.